

## Korrelation und Regression



Post hoc ergo propter hoc

In der Statistik bezieht sich der Begriff "Korrelation" auf die stochastische Abhängigkeit zweier Zufallsvariablen. So sind z.B. die Körpergröße und das Körpergewicht eines Menschen zweifellos korreliert, da große Leute tendenziell schwerer sind als kleine Leute. Wie fast alle Korrelationen ist aber auch dieser Zusammenhang nicht perfekt, da sich Menschen mit der gleichen Körpergröße erheblich in ihrem Körpergewicht unterscheiden können.

Leider hält die Einbeziehung von Korrelationen in die wissenschaftliche Inferenzbildung ein paar Fallstricke bereit. Zunächst einmal liefert eine Korrelation an sich noch keinen Beweis für Kausalität (denken Sie an die Störche und Babys). Zweitens werden beobachtete Korrelationen oft durch Messfehler abgeschwächt. Drittens können Korrelationen dann in die Irre führen, wenn der Zusammenhang zweier Merkmale in Wertebereiche extrapoliert wird, die durch die ursprüngliche Analyse gar nicht abgedeckt waren. Und schließlich besteht wie bei allen statistischen Tests auch bei der Bewertung von Korrelationen die Gefahr eines Typ-I-Fehlers, d.h. manche beobachtete Korrelationen können sich allein "durch Zufall" als statistisch signifikant erweisen.

Oftmals ist es möglich (und hoffentlich auch sinnvoll), die mathematische Beziehung zwischen zwei oder mehr Zufallsvariablen zu untersuchen. Im einfachsten Fall ist diese Beziehung linear. d.h. jede Erhöhung der einen Variablen um eine Einheit erhöht oder senkt die andere Variable um einen festen Betrag. In der Statistik wird das Modellieren des mathematischen Zusammenhangs zwischen Zufallsvariablen als "Regressionsanalyse" bezeichnet (die übrigens sehr viel mehr leisten kann, als das Anpassen einer Geraden an einen Datensatz).

## Mündliche Prüfungen

In einer mündlichen Examensprüfung wurden 100 Studierende von zwei Prüfern bewertet. Es ergaben sich 84 übereinstimmende Beurteilungen.

1. Prüfer

		1. Prüfer		
		bestanden	nicht bestanden	gesamt
2. Prüfer	bestanden	17	13	30
	nicht bestanden	3	67	70
	gesamt	20	80	100

Konkordanzrate (relative Übereinstimmung):  $\frac{17 + 67}{100} = 0.84$

Studenten dürfen von erfahrenen Professoren mit Fug und Recht erwarten, dass diese die Ergebnisse mündlicher Prüfungen konsistent und reproduzierbar bewerten. Mit anderen Worten sollten zwei Prüfer bei der Frage, ob ein Student ein Examen bestanden hat oder nicht, zum gleichen Ergebnis kommen. In diesem wie in vielen anderen Fällen ist es also sinnvoll, das Ausmaß der relativen Übereinstimmung zweier Entscheider durch die so genannte "Konkordanzrate" zu quantifizieren, d.h. durch den Anteil der konkordanten Entscheidungen an der Gesamtheit der Entscheidungen.

Im vorliegenden Beispiel beträgt die Konkordanzrate 0.84, oder 84% - aber ist dies wirklich ein nennenswertes Maß an Übereinstimmung? Nehmen wir einmal an, die Prüfer hätten während der Prüfungen aus Versehen vergessen, die Namen der Prüflinge zu notieren, und würden den Studenten die 30 bzw. 20 positiven Bewertungen nachträglich zulosen. Selbst in diesem Fall würde es ja eine gewisse Konkordanz geben, und diese "zufällig" zu erwartende Übereinstimmung sollte bei der Bewertung der tatsächlichen Konkordanz entsprechend in Rechnung gestellt werden.

## Konkordanzmaße

Cohens Kappa ( $\kappa$ )

$p_o$ : beobachtete Konkordanzrate

$p_e$ : "zufällig" erwartete Konkordanzrate

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

vollkommene Konkordanz ( $p_o=1$ )  $\longrightarrow$   $\kappa=1$

zufällige Übereinstimmung ( $p_o=p_e$ )  $\longrightarrow$   $\kappa=0$

Die am Ende der vorangegangenen Folie angedeutete Adjustierung führt zum beliebten "Kappa-Index" nach Cohen,  $\kappa$ , der oft auch als "zufallsbereinigtes" Übereinstimmungsmaß bezeichnet wird. Er setzt die überschüssige Konkordanz  $p_o - p_e$  zweier binärer Zufallsvariabler X und Y mit der zwischen X und Y bei stochastischer Unabhängigkeit erwarteten Diskordanz  $1 - p_e$  in Beziehung. Cohens  $\kappa$  gibt also an, welcher Anteil der zufällig erwarteten diskordanten Entscheidungen stattdessen konkordant ausgefallen sind.

Im Prinzip ist  $\kappa$  eine geeignete (wenn auch nicht besonders effiziente) Teststatistik für die stochastische Unabhängigkeit von X und Y. Allerdings ist die Frage nach der Unabhängigkeit der Entscheidungen zweier Bewerter in der Regel nicht besonders sinnvoll. In den meisten praktisch relevanten Situationen werden Bewerter *per definitionem* nicht unabhängig voneinander entscheiden, da sie ja ein und dieselbe Sache fair bewerten sollen.

Die Quantifizierung der Übereinstimmung zweier Bewerter mittels Cohens  $\kappa$  ist jedoch ebenfalls problematisch. Der in der Definition von  $\kappa$  enthaltene Anteil zufälliger Übereinstimmungen ist nur unter der unrealistischen Annahme der stochastischen Unabhängigkeit der Entscheidungen relevant. Es ist daher fraglich, ob dieser Anteil überhaupt zur "Bereinigung" von Übereinstimmungsmaßen geeignet ist. Ohne ein explizites Modell der Entscheidungsfindung ist unklar, wie sich der Zufall auf Entscheidungen auswirkt, und wie dies zu korrigieren wäre. Cohens  $\kappa$  sollte deswegen nicht unkritisch als Standard für die Bewertung von Übereinstimmungen angesehen werden. Darüber hinaus sind Tabellen, die eine Klassifizierung möglicher  $\kappa$ -Werte als "gut", "mittelmäßig", "schlecht" usw. anregen, aus theoretischer Sicht unangemessen und sollten daher auch nicht benutzt werden.

## Mündliche Prüfungen

1. Prüfer

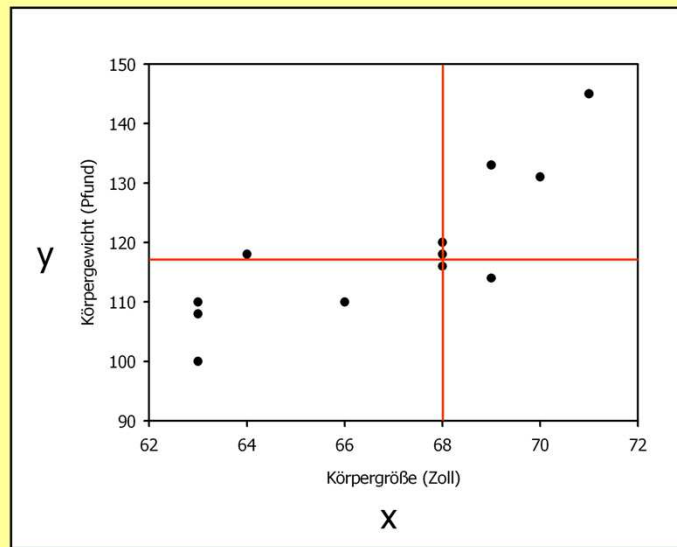
		bestanden	nicht bestanden	gesamt
2. Prüfer	bestanden	17	13	30
	nicht bestanden	3	67	70
	gesamt	20	80	100

$$p_e = \frac{30}{100} \cdot \frac{20}{100} + \frac{70}{100} \cdot \frac{80}{100} = 0.62$$

$$\rightarrow \kappa = \frac{0.84 - 0.62}{1 - 0.62} = 0.58$$

Im vorliegenden Beispiel würde man selbst bei Zulosen der positiven Bewertungen immerhin für 62 der 100 Prüflinge eine Übereinstimmung der Entscheidungen erwarten. Das ist weniger als die beobachteten 84, aber der Unterschied ist nicht besonders groß. In der Tat beträgt  $\kappa$  nur 0.58, d.h. nur 58% der bei Zulosen erwarteten diskordanten Entscheidungen waren stattdessen konkordant. Die beiden Prüfer hätten also noch deutlich besser abschneiden können.

### Miss America 1984 - 2002



Im Prinzip lässt sich die Stärke des Zusammenhangs zweier quantitativer Zufallsvariablen auch mit Hilfe eines Konkordanzmaßes für nominale Daten (wie z.B. Cohens  $\kappa$ ) quantifizieren. Dies setzt natürlich voraus, dass die Daten zuvor in ein passendes Format gebracht wurden. Eine beliebte Methode zur Transformation quantitativer in qualitative Daten ist die Teilung entlang des Medians, d.h. die ursprünglichen Daten werden durch ihre Klassifizierung relativ zum Median ("größer", "kleiner") ersetzt.

Ein Grund dafür, weshalb solche Dichotomisierungen bei medizinischen Wissenschaftlern sehr beliebt sind, besteht in dem bei Medizinern ohnehin weit verbreiteten Denken in Gruppen oder Klassen. Manche Leute sind "hypertensiv", andere sind "normotensiv", oder Personen werden als "schnelle" bzw. "langsame" Metabolisierer klassifiziert. Es gibt in der Tat eine Reihe guter Gründe, sich diese Sichtweise anzueignen (einschließlich der intellektuellen Klarheit und der einfacheren Kommunizierbarkeit), statt sich Personen entlang einer eindimensionalen Skala von Blutdruckwerten oder Metabolisierungsraten aufgereiht vorzustellen.

### Miss America 1984 - 2002

		Körpergröße		
		<68"	≥68"	gesamt
Körpergewicht	<117 lbs	4	2	6
	≥117 lbs	1	5	6
	gesamt	5	7	12

$$p_o = \frac{4 + 5}{12} = 0.75 \quad p_e = \frac{6}{12} \cdot \frac{5}{12} + \frac{6}{12} \cdot \frac{7}{12} = 0.50$$

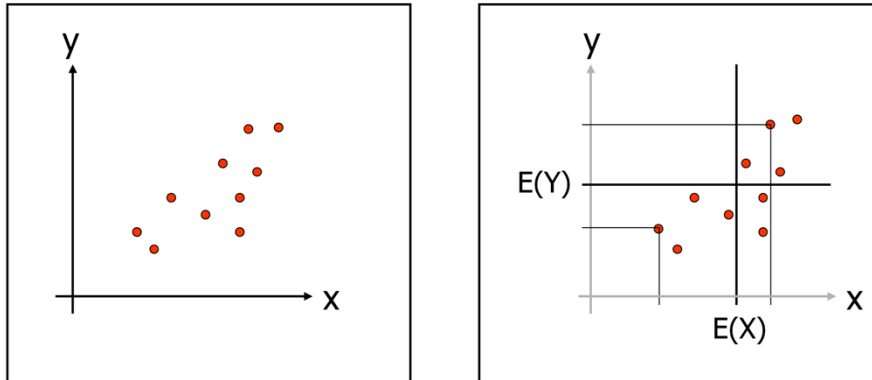
$$\rightarrow \kappa = \frac{0.75 - 0.50}{1 - 0.50} = 0.50$$

Die nähere Betrachtung der dichotomisierten Miss America Größen- und Gewichtsdaten bestätigt die praktische Erfahrung, dass große Leute tendenziell schwerer sind als kleine Leute, was offensichtlich auch für Schönheitsköniginnen gilt. Die Konkordanz der Klassifizierungen als "groß" bzw. "klein" und "schwer" bzw. "leicht" ist ausgeprägt, aber nicht perfekt. Insgesamt stimmten 9 von 12 Klassifizierungen überein, was einem Überschuss von 3 entspricht, d.h. 50% der 6 bei Unabhängigkeit erwarteten diskordanten Klassifizierungen waren stattdessen konkordant. Daraus ergibt sich  $\kappa=0.5$ . Beachten Sie jedoch, dass Cohens  $\kappa$  in diesem Kontext ziemlich bedeutungslos ist, da die stochastische Unabhängigkeit von Körpergröße und -gewicht physiologisch unmöglich ist.

Auch wenn Dichotomisierungen einfach und einleuchtend seien mögen, so bedeutet die Transformation quantitativer in qualitative Daten in der Regel einen erheblichen Verlust an Information und damit verbunden an Power für den Nachweis tatsächlicher Assoziationen. Aus diesem Grund warnen die meisten Statistiker vor Dichotomisierungen und empfehlen stattdessen die Analyse der Originaldaten, wann immer dies möglich ist.

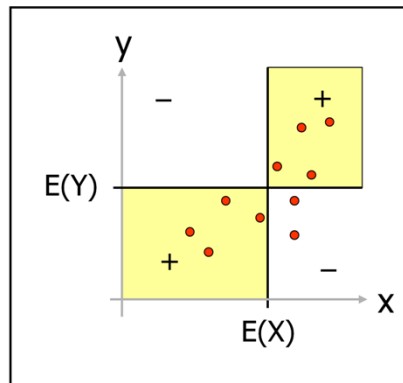
## Korrelation

X und Y zwei Zufallsvariable  
10 Realisierungen



Im vorliegenden Beispiel legt schon der Augenschein von 10 gemeinsamen Realisierungen  $(x,y)$  zweier quantitativer Zufallsvariablen X und Y den Schluss nahe, dass diese beiden Zufallsvariablen nicht stochastisch unabhängig voneinander sind. Je größer die X-Werte, umso größer sind die zugehörigen Y-Werte. Mit anderen Worten gehen große positive (bzw. negative) Abweichungen zwischen X und seinem Erwartungswert  $E(X)$  mit großen positiven (bzw. negativen) Abweichungen zwischen Y und  $E(Y)$  einher.

## Kovarianz $\text{Cov}(X,Y)$



$x > E(X), y > E(Y)$  oder  $x < E(X), y < E(Y)$   $\rightarrow [x - E(X)] \cdot [y - E(Y)] > 0$

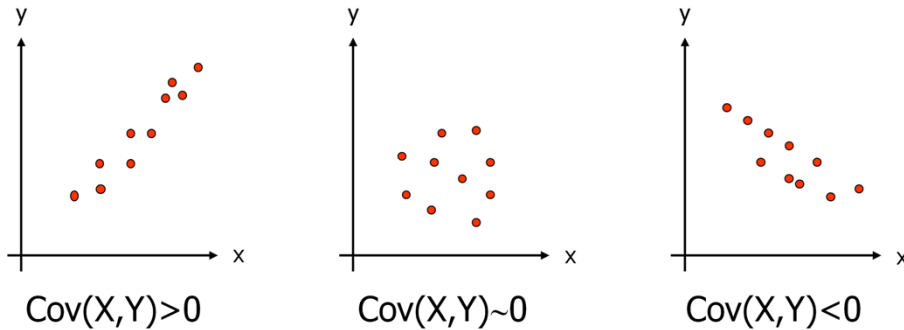
$x > E(X), y < E(Y)$  oder  $x < E(X), y > E(Y)$   $\rightarrow [x - E(X)] \cdot [y - E(Y)] < 0$

$$\text{Cov}(X,Y) = E([X - E(X)] \cdot [Y - E(Y)])$$

Da die meisten gemeinsamen Realisierungen  $(x,y)$  von  $X$  und  $Y$  in die mit "+" markierten Quadranten fallen, ist auch das Produkt aus  $[x - E(X)]$  und  $[y - E(Y)]$  in den meisten Fällen positiv. Beachten Sie dabei, dass jeder Wert von  $[x - E(X)] \cdot [y - E(Y)]$  selbst wieder die Realisierung einer zusammengesetzten Zufallsvariablen ist, nämlich des Produkts  $[X - E(X)] \cdot [Y - E(Y)]$ . Der Erwartungswert dieser Zufallsvariablen heißt die "Kovarianz" von  $X$  und  $Y$ , abgekürzt  $\text{Cov}(X,Y)$ . Wenn  $\text{Cov}(X,Y) = 0$  gilt, so bezeichnet man die beiden Zufallsvariablen  $X$  und  $Y$  als "unkorreliert". Gemessen an der Lage der 10 gemeinsamen Realisierungen  $(x,y)$  scheint die Kovarianz von  $X$  und  $Y$  in unserem Beispiel positiv zu sein, d.h.  $X$  und  $Y$  sind positiv korreliert. Würden stattdessen die meisten Beobachtungspaare in die mit einem Minuszeichen markierten Quadranten fallen, so würde dies den Schluss  $\text{Cov}(X,Y) < 0$  nahe legen.



## Kovarianz $\text{Cov}(X,Y)$



Wenn in jeden der vier auf der vorangegangenen Folie skizzierten Quadranten die gleiche Anzahl paarweiser Realisierungen  $(x,y)$  fallen würde, so wären ungefähr gleich viele Produkte  $[x-E(X)] \cdot [y-E(Y)]$  positiv wie negativ. Im Durchschnitt würde  $[x-E(X)] \cdot [y-E(Y)]$  dann also null betragen. Wenn wir uns nun erinnern, dass der Erwartungswert einer Zufallsvariablen dem arithmetischen Mittel einer großen Anzahl von Realisierungen entspricht, dann ergibt sich für das Verteilungsmuster in der Mitte der vorliegenden Folie die Vermutung einer Kovarianz dicht bei null, d.h.  $\text{Cov}(X,Y) \sim 0$ .

## Kovarianz $\text{Cov}(X,Y)$

Es seien  $(X,Y)$  ein Paar von Zufallsvariablen und  $(x_1, y_1), \dots, (x_n, y_n)$  eine Stichprobe aus  $n$  unabhängigen Realisierungen von  $(X,Y)$ .

Ein "guter" Schätzer (d.h. unverzerrt, konsistent, meistens effizient) der Kovarianz  $\text{Cov}(X,Y)$  ist die empirische Kovarianz.

$$s_{XY} = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Miss America 1984 - 2002

$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$
63"	100 lbs	-3.8"	-18.6 lbs
68"	120 lbs	1.2"	1.4 lbs
69"	114 lbs	2.2"	-4.6 lbs
68"	116 lbs	1.2"	-2.6 lbs
70"	131 lbs	3.2"	12.4 lbs
63"	108 lbs	-3.8"	-10.6 lbs
68"	118 lbs	1.2"	-0.6 lbs
66"	110 lbs	-0.8"	-8.6 lbs
71"	145 lbs	4.2"	26.4 lbs
69"	133 lbs	2.2"	14.4 lbs
64"	118 lbs	-2.8"	-0.6 lbs
63"	110 lbs	-3.8"	-8.6 lbs

$$\bar{x} = 66.8 \text{ "}$$

$$\bar{y} = 118.6 \text{ lbs}$$

$$s_{xy} = 29.3 \text{ "}\cdot\text{lbs}$$

Beachten Sie, dass die Kovarianz nicht dimensionslos ist.

Miss America 1984 - 2002

$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$
1.60 m	45.4 kg	-0.10 m	-8.4 kg
1.73 m	54.4 kg	0.03 m	0.6 kg
1.75 m	51.7 kg	0.05 m	-2.1 kg
1.73 m	52.6 kg	0.03 m	-1.2 kg
1.78 m	59.4 kg	0.08 m	5.6 kg
1.60 m	49.0 kg	-0.10 m	-4.8 kg
1.73 m	53.5 kg	0.03 m	-0.3 kg
1.68 m	49.9 kg	-0,02 m	-3.9 kg
1.80 m	65.8 kg	0.10 m	12.0 kg
1.75 m	60.3 kg	0.05 m	6.5 kg
1.63 m	53.5 kg	-0.07 m	-0.3 kg
1.60 m	49.9 kg	-0.01 m	-3.9 kg

$$\begin{aligned}\bar{x} &= 1.70 \text{ m} \\ \bar{y} &= 53.8 \text{ kg} \\ s_{xy} &= 0.33 \text{ m}\cdot\text{kg}\end{aligned}$$

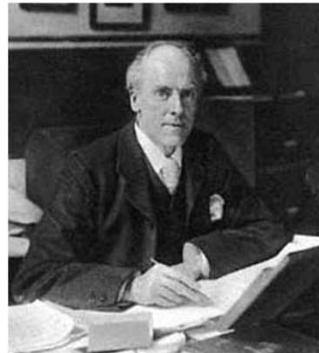
Die Daten auf dieser Folie sind die gleichen wie auf der vorangegangenen, allerdings wurden die Werte von Zoll in Meter und von imperialen Pfund in Kilogramm transformiert. Als Konsequenz dieser Reskalierung hat sich die geschätzte Kovarianz der beiden Zufallsvariablen von 29.3 Zoll-Pfund in 0.33 Meter-Kilogramm geändert. Abgesehen von seinem Vorzeichen stellt die Kovarianz also kein besonders gutes Assoziationsmaß dar. Ein Punktdiagramm zur Visualisierung des Zusammenhangs zwischen Körpergewicht und -größe würde schließlich auch immer gleich aussehen, unabhängig von der Skalierung. Von einem guten Assoziationsmaß kann man die gleiche Eigenschaft erwarten; es sollte daher dimensionslos sein.

## Pearson-Korrelationskoeffizient

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

$-1 \leq r_{XY} \leq +1$  ist dimensionslos

$$\hat{r}_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$



Karl Pearson  
(1857-1936)

Karl Pearson wurde am 27. März 1857 in eine Londoner Familie des gehobenen Mittelstands hineingeboren. Sein Vater war Rechtsanwalt. Pearson erhielt zunächst Privatunterricht an der University College School und nahm später ein Mathematikstudium am King's College in Cambridge auf. Im Jahr 1884 wurde Karl Pearson zum Professor für Angewandte Mathematik und Mechanik am University College, London, berufen. Seine Vorlesungen über Wahrscheinlichkeit und über die Philosophie des Zufalls bildeten die Grundlage seiner späteren biometrischen und statistischen Arbeiten. Durch Vermittlung von Walter Weldon, einem Zoologieprofessor am University College, lernte Pearson den Vater der Forschung über die interindividuelle Variabilität von Lebewesen kennen: Sir Francis Galton. Durch seinen Kontakt mit Galton entwickelte Pearson ein mathematisch-methodisches Interesse an Fragen der Erbllichkeit und der Evolution. Sein dabei ebenfalls an den Tag gelegtes biologisches Interesse war wegbereitend für die Entwicklung der heute als "Biostatistik" bezeichneten Wissenschaft.

Karl Pearson ist am besten bekannt für das statistische Maß, das auch seinen Namen trägt, den "Pearson-Korrelationskoeffizienten". Neben der Zusammenhangsanalyse konzipierte Pearson (mit finanzieller und intellektueller Unterstützung durch Galton) auch noch den  $\chi^2$ -Test. Diese beiden Methoden waren grundlegend für die Entwicklung der multivariaten Statistik als neue mathematische Disziplin. Nebenbei bemerkte war Karl Pearson der Vater von Egon Sharpe Pearson, dem großen britischen Mathematiker, der in den 1930er Jahren gemeinsam mit Jerzy Neyman die Prinzipien des statistischen Hypothesentests entwickelte. Karl Pearson starb am 27. April 1936 in London.

Miss America 1984 - 2002

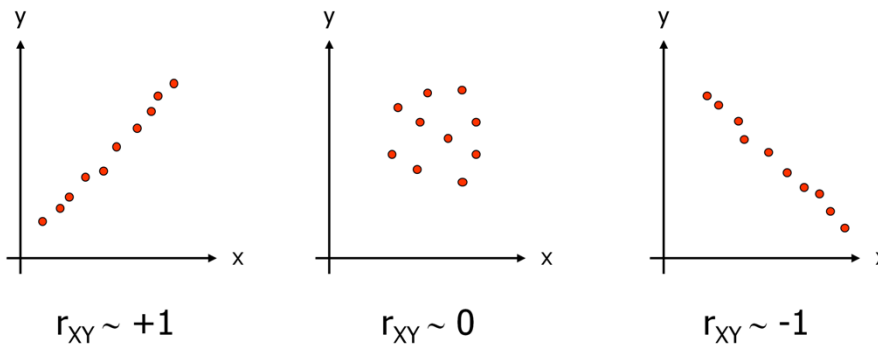
63"	100 lbs	1.60 m	45.4 kg
68"	120 lbs	1.73 m	54.4 kg
69"	114 lbs	1.75 m	51.7 kg
68"	116 lbs	1.73 m	52.6 kg
70"	131 lbs	1.78 m	59.4 kg

63"	$s_{XY} = 29.3 \text{ "}\cdot\text{lbs}$ $s_X = 2.9 \text{ "}$ $s_Y = 12.4 \text{ lbs}$ $\hat{r}_{XY} = +0.81$	60 m
68"		73 m
66"		68 m
71"		80 m
69"		75 m
64"		63 m
63"	60 m	

$s_{XY} = 0.33 \text{ m}\cdot\text{kg}$ $s_X = 0.073 \text{ m}$ $s_Y = 5.6 \text{ kg}$ $\hat{r}_{XY} = +0.81$	60 m
	73 m
	68 m
	80 m
	75 m
	63 m
60 m	

Der geschätzte Pearson-Korrelationskoeffizient der Größen- und Gewichtsdaten der Miss America beträgt 0.81, unabhängig von der Skalierung der zu Grunde liegenden Messwerte.

## Pearson-Korrelationskoeffizient



$r_{XY}$  misst die Stärke und Richtung des **linearen Zusammenhangs** zwischen X und Y.

Die möglichen Werte des Pearson-Korrelationskoeffizienten  $r_{XY}$  liegen zwischen -1 und +1. Der Absolutbetrag von  $r_{XY}$  gibt die Stärke des linearen Zusammenhangs von X und Y an, d.h. wie gut sich ein Streudiagramm gemeinsamer Realisierungen  $(x,y)$  durch eine Gerade annähern lässt. Werte dicht bei +1 oder -1 weisen auf einen starken linearen Zusammenhang zwischen X und Y hin. Das Vorzeichen von  $r_{XY}$  wiederum spiegelt die Richtung des linearen Zusammenhangs wider. Wenn  $r_{XY} > 0$  gilt, liegt ein positiver Zusammenhang vor, bei  $r_{XY} < 0$  ein negativer.

## Pearson-Korrelationskoeffizient

### Interpretation

Besteht zwischen zwei Zufallsvariablen X und Y ein linearer Zusammenhang, so bezeichnet man diesen als

perfekt	↔	$ r_{XY}  = 1.00$
stark	↔	$0.75 \leq  r_{XY}  < 1.00$
moderat	↔	$0.50 \leq  r_{XY}  < 0.75$
schwach	↔	$0.25 \leq  r_{XY}  < 0.50$

Gilt  $|r_{XY}| < 0.25$ , so spricht man nicht mehr von einem linearen Zusammenhang zwischen X und Y.



## Pearson-Korrelationskoeffizient Signifikanztest

**Zufallsvariable**  $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$  alle unbekannt

**Hypothesen**  $H_0 : r_{XY} = 0 \quad H_A : r_{XY} \neq 0$  (zweiseitig)  
 $H_0 : r_{XY} \leq 0 \quad H_A : r_{XY} > 0$  (einseitig)

**Teststatistik**  $T = \hat{r}_{XY} \cdot \sqrt{\frac{n-2}{1-\hat{r}_{XY}^2}}$

**Ablehnungsbereich**  $T \leq t_{\alpha/2, n-2}$  oder  $T \geq t_{1-\alpha/2, n-2}$  (zweiseitig)  
 $T \geq t_{1-\alpha, n-2}$  (einseitig)

Unter der Annahme, dass X und Y beide einer Normalverteilung folgen, kann die Schätzung von  $r_{XY}$  aus einer Stichprobe auch zur Bewertung der statistischen Signifikanz einer beobachteten Korrelation verwendet werden. Diese Statistik beträgt z.B. für die Miss America

$$t = 0.81 \cdot \sqrt{\frac{12-2}{1-0.81^2}} = 4.368,$$

wodurch der kritische Wert  $t_{0.975, 10} = 2.228$  bei weitem übertroffen wird. Wie nicht anders zu erwarten, ist die beobachtete Korrelation zwischen der Körpergröße und dem Körpergewicht der Schönheitsköniginnen also statistisch signifikant.

Beachten Sie, dass der auf dieser Folie vorgestellte Test nur dann valide ist, wenn es sich bei den Stichprobenwerten x und y auch tatsächlich um Realisierungen zweier normalverteilter Zufallsvariabler X und Y handelt. Dies ist z.B. dann nicht der Fall, wenn die Einflussgröße X systematisch gemäß eines experimentellen Plans oder Designs variiert wurde (also z.B. um einen festen Betrag erhöht oder reduziert wurde). Darüber hinaus müssen X und Y noch eine Reihe weiterer Voraussetzungen erfüllen, insbesondere dass ihre gemeinsame Verteilung im zweidimensionalen Raum ebenfalls wieder normal ist. All diese Einschränkungen haben zur Folge, dass die praktische Anwendbarkeit des dargestellten Tests im Allgemeinen problematisch ist.

## Regressionsmodelle

Regressionsmodelle dienen dazu,  
eine Variable aus einer oder mehr  
anderen Variablen vorherzusagen.

Wissenschaftler nutzen Regressionsmodelle für  
Aussagen über ein Ereignis, wenn die dafür  
notwendigen Informationen billiger zu bekommen sind  
als Informationen über das Ereignis selbst, oder wenn  
das Ereignis erst in der Zukunft eintritt.

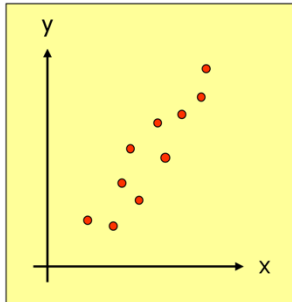
<http://www.psychstat.smsu.edu/introbook/sbk16.htm>

Die Regressionsanalyse ist eines der am häufigsten angewandten statistischen Verfahren in den Natur- und Sozialwissenschaften und in der medizinischen Forschung. Ihr Hauptziel ist die Charakterisierung des Zusammenhangs zwischen einer zufälligen Zielgröße (manchmal auch als "abhängige Variable" bezeichnet) und einer oder mehreren Einflussgrößen (auch "unabhängige" oder "erklärende Variable" genannt). Die Werte der Einflussgrößen können entweder fest vorgegeben oder zufällig sein. Ziel der Regressionsanalyse ist die Ermittlung einer mathematischen Funktion (eines "Regressionsmodells"), die sich am besten einer Stichprobe aus gemeinsamen Realisierungen der Einfluss- und Zielgrößen anpasst.

Regressionsanalysen sollten nur durchgeführt werden, wenn das dabei abgeleitete Modell auch irgendwelche Konsequenzen hat, z.B. in Form eines besseren Verständnisses der den Daten zugrunde liegenden Mechanismen oder zur praktischen Vorhersage zukünftiger Ereignisse. Viel zu oft führen Wissenschaftler aber eine Regressionsanalyse einzig und allein mit dem Ziel durch, eine Gerade durch ihre Beobachtungsdaten zeichnen zu können.

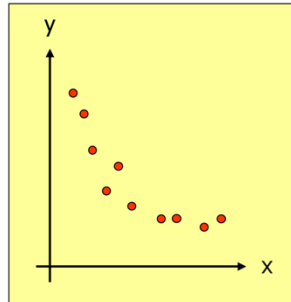
## Regressionsmodelle

lineare  
Regression



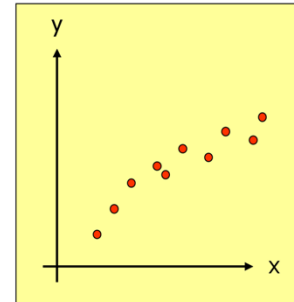
$$y = a + b \cdot x$$

exponentielle  
Regression



$$y = a \cdot e^{-b \cdot x}$$

logarithmische  
Regression



$$y = a + b \cdot \log(x)$$

Das einfachste Regressionsmodell ist das lineare, d.h.  $y = a + b \cdot x$ , welches zwei Variable mittels einer Geraden in Beziehung zueinander setzt. Die lineare Regression stellt jedoch insoweit eine Besonderheit dar, als sich der Achsenabschnitt und die Steigung der Geraden wegen der Einfachheit des mathematischen Zusammenhangs zwischen den Variablen nötigenfalls auch zu Fuß ausrechnen lassen. Die nicht-lineare Regression verwendet demgegenüber spezifische nicht-lineare Funktionen, um Zielgrößen aus Einflussgrößen vorherzusagen. Die Art des Zusammenhangs, d.h. die Form der im konkreten Fall in Frage kommenden Regressionsfunktionen, erschließt sich entweder aus theoretischen Überlegungen oder resultiert aus der explorativen Analyse unabhängiger Daten.

Bei der nicht-linearen Regression können Probleme auftreten, die es bei der linearen Regression in der Form nicht gibt. So kann sich z.B. die Schätzung der Regressionsparameter als sehr schwierig gestalten, und die verwendeten Schätzer sind unter Umständen nicht unverzerrt. Daher versuchen Wissenschaftler oft, ihre Daten so zu transformieren (z.B. durch Logarithmieren), dass aus der nicht-linearen Beziehung zwischen den Originaldaten ein linearer Zusammenhang zwischen den transformierten Daten entsteht. Diese Praxis kann allerdings zu Interpretationsproblemen führen, wenn das lineare Regressionsmodell zur Modellierung der Originaldaten einfach nur zurück transformiert wird.

## Lineare Regression

Y: Zielgröße

X: Einflussgröße

E: Zufallsfehler

Für E wird im Allgemeinen eine  $N(0, \sigma^2)$ -Verteilung mit unbekanntem  $\sigma^2$  unterstellt.

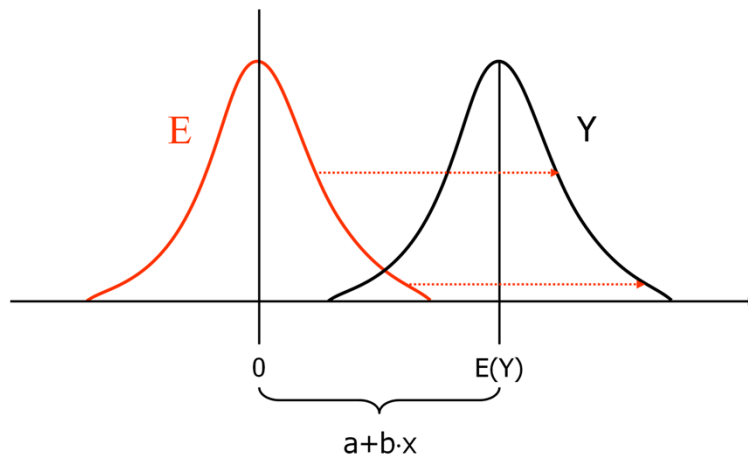
$$Y = a + b \cdot x + E$$

Steigung b heißt "Regressionskoeffizient"

Wie man sieht, enthält die mathematisch korrekte Form des linearen Regressionsmodells zwei Großbuchstaben, Y und E (der große griechische Buchstabe "Epsilon"), die jeweils eine Zufallsvariable bezeichnen. Für den Wert x der Einflussgröße X wurde ein kleiner Buchstabe gewählt, da seine Rolle im Regressionsmodell deterministisch ist. Wenn Y in der Realität "erzeugt" (d.h. realisiert) wird, ist x bekannt, und die Zufälligkeit von Y resultiert einzig und allein aus dem Zufallsfehler E. Mit anderen Worten besteht der Unterschied zweier Realisierungen von Y bei gleichem Wert der Einflussgröße X nur im Beitrag des Zufalls. Der X-Wert dient den beiden Y-Werten als gemeinsamer Ausgangspunkt.

Interessanterweise erfüllen zwei normalverteilte Zufallsvariable X und Y immer automatisch ein lineares Regressionsmodell. Wenn also der Zusammenhang zwischen X und Y ganz offensichtlich nicht linear ist (was sich oft schon den Daten selbst entnehmen lässt), dann können nicht beide Zufallsvariable gleichzeitig einer Normalverteilung folgen. In diesem Fall würde es auch keinen Sinn machen, die Signifikanz von  $r_{XY} \neq 0$  mit Hilfe des auf einer der vorangegangenen Folien eingeführten Tests zu bewerten. Dieser Test wäre nicht valide.

## Lineare Regression



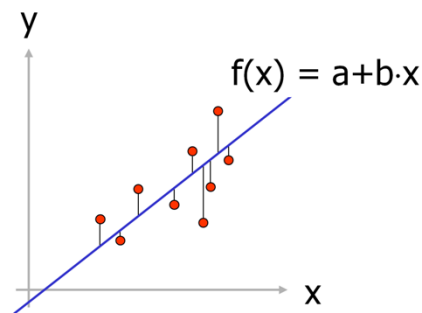
Für einen gegebenen Wert  $x$  von  $X$  beträgt der Erwartungswert von  $Y$

$$E(Y) = E(a + b \cdot x + E) = a + b \cdot x + E(E) = a + b \cdot x,$$

weil  $E(E)$  üblicherweise als null vorausgesetzt wird. Somit ist die Verteilung von  $Y$  die gleiche wie die von  $E$ , allerdings verschoben um den deterministischen Anteil  $a+b \cdot x$  des Regressionsmodells, d.h. um den Erwartungswert von  $Y$ .

## Kleinste-Quadrate-Prinzip (lineare Regression)

Wähle a und b so, dass  $\sum_{i=1}^n (a + b \cdot x_i - y_i)^2$  minimal ist!



Das Kleinste-Quadrate-Prinzip stellt sicher, dass die Schätzer von a und b bei Gültigkeit eines linearen Modells unverzerrt, konsistent und effizient sind.

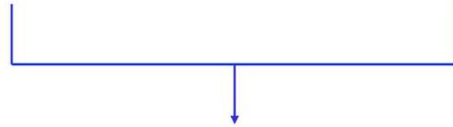
Ziel jeder Regressionsanalyse ist die Identifizierung einer mathematischen Funktion, durch die sich die Stichprobendaten am besten annähern lassen. Eine gute Regressionsfunktion sollte daher einen "best fit" in dem Sinne liefern, dass sich die Werte der Zielgröße möglichst präzise aus den Werten der Einflussgrößen vorhersagen lassen. Ein intuitiv einleuchtendes Kriterium für die Entscheidung über den "best fit" sind die senkrechten Abstände (auch "Residuen" genannt) zwischen den tatsächlichen Datenpunkten und den hypothetischen Datenpunkten, die durch eine Anwendung der Regressionsfunktion auf die Werte der Einflussgrößen entstehen würden. Demgemäß versuchen Regressionsanalysen bei der Suche nach dem "best fit", die Quadratsumme dieser Residuen zu minimieren. Die Verwendung von Quadraten sorgt dafür, dass mehr Gewicht auf große Diskrepanzen zwischen vorhergesagten und tatsächlichen Werten gelegt wird als auf kleine Unterschiede.

Damit das hier dargestellte Kleinste-Quadrate-Prinzip auch zu sinnvollen Regressionsmodellen führt, muss es sich bei den individuellen Messfehlern  $\varepsilon_i$  um Realisierungen unkorrelierter Zufallsvariabler  $E_i$  handeln, die wiederum selbst  $E(E_i)=0$  erfüllen und alle die gleiche Varianz haben. Unter diesen Umständen folgt aus dem so genannten "Gauß-Markov-Theorem", dass die Kleinste-Quadrate-Schätzer der Regressionskoeffizienten tatsächlich unverzerrt sind. Sind die Fehler zusätzlich normalverteilt, so sind die Kleinste-Quadrate-Schätzer von a und b auch konsistent und effizient.

## Lineare Regression

$$\sum_{i=1}^n (a + b \cdot x_i - y_i)^2 \text{ minimal!}$$

$$\frac{\delta}{\delta a} \sum_{i=1}^n (a + b \cdot x_i - y_i)^2 = 0 \quad \frac{\delta}{\delta b} \sum_{i=1}^n (a + b \cdot x_i - y_i)^2 = 0$$



$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{und} \quad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

Für lineare Modelle lassen sich die Kleinste-Quadrate-Schätzer der Regressionskoeffizienten noch analytisch ermitteln, d.h. es gibt explizite Formeln, mit denen sich der Achsenabschnitt  $a$  und die Steigung  $b$  aus den vorliegenden Daten ausrechnen lassen. Demgegenüber muss die Quadratsumme der Residuen bei nicht-linearen Modellen in der Regel mit Hilfe iterativer Algorithmen im Computer minimiert werden.

Miss America 1984 - 2002

$x_i - \bar{x}$	$y_i - \bar{y}$
-3.8"	-18.6 lbs
1.2"	1.4 lbs
2.2"	-4.6 lbs
1.2"	-2.6 lbs
3.2"	12.4 lbs
-3.8"	-10.6 lbs
1.2"	-0.6 lbs
-0.8"	-8.6 lbs
4.2"	26.4 lbs
2.2"	14.4 lbs
-2.8"	-0.6 lbs
-3.8"	-8.6 lbs

$$\bar{x} = 66.8 \text{ "}$$

$$\bar{y} = 118.6 \text{ lbs}$$

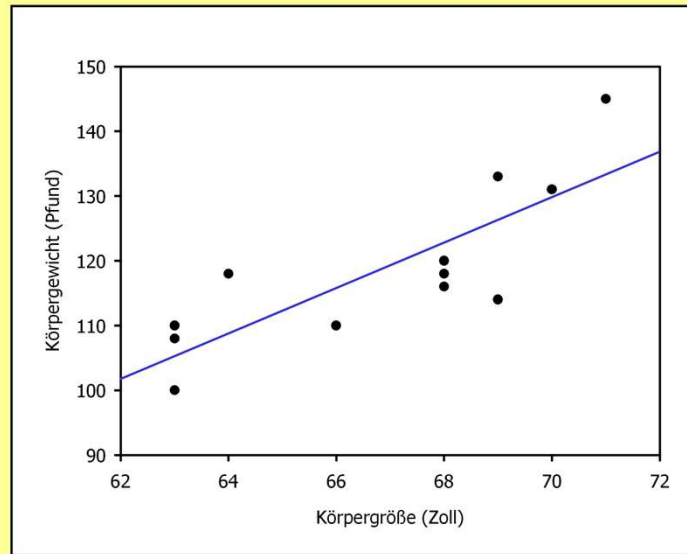
$$\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 322.16$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 93.68$$

$$\hat{a} = -111.29 \quad \hat{b} = 3.44$$



### Miss America 1984 - 2002



Eine Regressionsfunktion sollte nur auf den Wertebereich angewandt werden, auf den sich auch die ursprüngliche Regressionsanalyse gestützt hat. Die Miss America Daten werden z.B. am besten durch das lineare Modell  $y = -111.29 + 3.44 \cdot x$  angenähert. Es ist aber sehr unwahrscheinlich, dass das gleiche Modell auch zu Miss Lilliput Daten passen würde, da es für ein 32.35 Zoll großes Mädchen ein Körpergewicht von  $y = 0.00$  Pfund vorhersagt!

## Bestimmtheitsmaß

Es seien  $(X,Y)$  ein Paar von Zufallsvariablen und  $(x_1,y_1), \dots, (x_n,y_n)$  eine Stichprobe aus  $n$  unabhängigen Realisierungen von  $(X,Y)$ .

$$R^2 = \frac{\sum_{i=1}^n [f(x_i) - \bar{y}]^2}{\sum_{i=1}^n [y_i - \bar{y}]^2}$$

$R^2$  bezeichnet den Anteil an der Variation der Zielgröße, der durch das Regressionsmodell  $f(x)$  erklärbar ist.

Zur Quantifizierung, wie gut ein Regressionsmodell einen gegebenen Datensatz beschreibt, wird oft das so genannte "Bestimmtheitsmaß"  $R^2$  verwendet. Da  $R^2$  die Realisierung einer (zugegebenermaßen komplizierten) Zufallsvariablen ist, kann im Kontext wissenschaftlicher Studien auch die Angabe eines Konfidenzintervalls für den Erwartungswert von  $R^2$  sinnvoll sein.

## Bestimmtheitsmaß

Für die einfache lineare Regression gilt

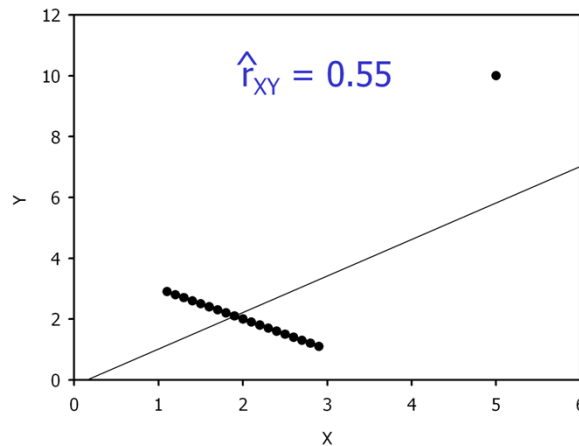
$$R^2 = \hat{r}_{XY}^2$$

d.h.  $\hat{r}_{XY}^2$  bezeichnet den Anteil an der beobachteten Variation der Zielgröße Y, der durch die Variation der Werte der Einflussgröße X und die Gerade  $f(x)=a+b \cdot x$  erklärbar ist.

Für ein einfaches lineares Regressionsmodell zweier Zufallsvariablen X und Y entspricht  $R^2$  dem Quadrat des Pearson-Korrelationskoeffizienten, d.h.  $R^2 = \hat{r}_{XY}^2$ . Da für die Miss America Daten  $\hat{r}_{XY}=0.81$  gilt, bedeutet dies, dass mit Hilfe eines linearen Modells 66% ( $0.81^2=0.66$ ) der Variation im Körpergewicht der Schönheitsköniginnen durch deren unterschiedliche Körpergröße erklärbar ist.

## Pearson-Korrelationskoeffizient

$\hat{r}_{XY}$  ist anfällig gegen Ausreißer!



Ausreißer sind atypische und (*per definitionem*) seltene Beobachtungen. Wegen der Art und Weise, wie eine Regressionsgerade zustande kommt, nämlich durch Minimierung der Quadratsumme der Residuen, haben Ausreißer einen erheblichen Einfluss auf die Steigung der Gerade und damit den Pearson-Korrelationskoeffizienten. Wenige Ausreißer können dessen Schätzung erheblich verändern, wie auf der vorliegenden Folie eindrucksvoll dargestellt ist. In unserem Beispiel reicht sogar ein einziger Ausreißer, um einen geschätzten Korrelationskoeffizienten von ursprünglich -1 auf einen relativ hohen Wert von 0.55 zu drücken. Es erübrigt sich fast der Hinweis, dass man wichtige Schlussfolgerungen aus Daten niemals allein auf dem geschätzten Korrelationskoeffizienten basieren lassen sollte, ohne zuvor ein Streudiagramm der Daten in Augenschein genommen zu haben.

## Spearman-Rang-Korrelationskoeffizient

Es seien  $rg[x_i]$  und  $rg[y_i]$  die Ränge von  $x_i$  bzw.  $y_i$ , wobei Bindungen durchschnittliche Ränge zugewiesen werden.

Ersetze  $x_i$  durch  $rg[x_i]$  und  $y_i$  durch  $rg[y_i]$ .

Berechne den Pearson-Korrelationskoeffizient für  $rg[x_i]$  und  $rg[y_i]$ .



Charles E. Spearman  
(1863-1945)

Ein alternatives Maß für die Stärke des Zusammenhangs zwischen zwei Zufallsvariablen wurde 1904 von dem englischen Psychologen Charles E. Spearman vorgeschlagen. Der "Spearman-Rang-Korrelationskoeffizient" erlaubt die Bewertung der Stärke des monotonen (nicht des linearen) Zusammenhangs, insbesondere in Situationen, in denen die Verwendung des Pearson-Korrelationskoeffizienten auf Grund der Verteilung der Daten irreführend ist. Da der Spearman Korrelationskoeffizient nicht auf den eigentlichen Daten basiert, sondern nur auf deren Rängen, ist er weniger anfällig gegen Ausreißer als der Pearson-Korrelationskoeffizient.

Spearman wurde am 10. September 1863 in London geboren. Er ließ sich in Deutschland ausbilden und promovierte 1904 nach Studien in Leipzig, Würzburg und Göttingen an der Leipziger Universität. Spearman verbrachte ein Viertel seines langen, 82 Jahre umfassenden Lebens in der britischen Armee (ein Grund dafür, weshalb er seine Doktorarbeit erst mit 41 Jahren abschloss). Zwischen 1907 und 1931 hatte er mehrere Lehrstühle am University College London inne, zunächst für Experimentelle Psychologie, später für Philosophie des Denkens und der Logik. Charles Spearman hat viele Theorien und Entdeckungen zur Grundlage des psychologischen Testens beigesteuert. Interessanterweise gewannen seine statistischen Arbeiten jedoch nie die Anerkennung seines berühmten Kollegen am University College, Karl Pearson; stattdessen bestand eine lang andauernde Fehde zwischen den beiden.

## Spearman-Rang-Korrelationskoeffizient

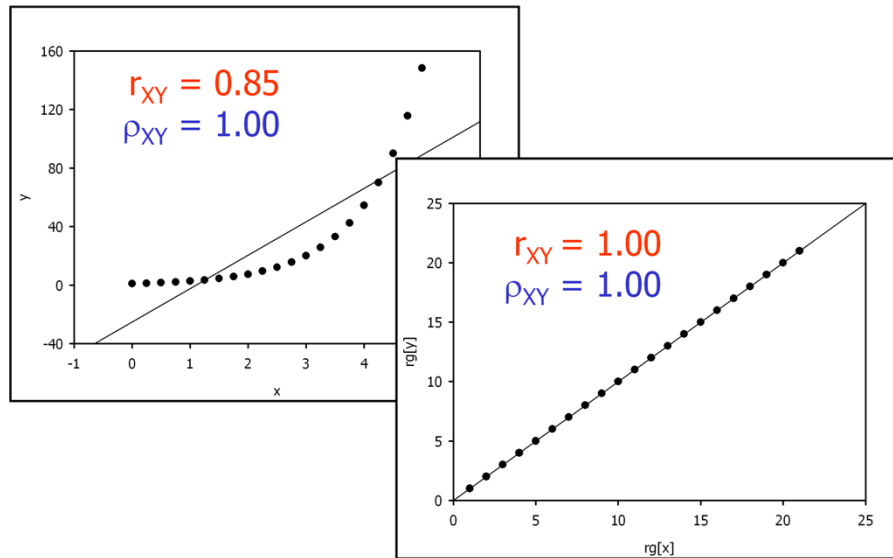
$$\hat{\rho}_{XY} = 1 - \frac{6 \cdot \sum_{i=1}^n \overbrace{(\text{rg}[x_i] - \text{rg}[y_i])^2}^{d_i^2}}{n \cdot (n^2 - 1)}$$

$-1 \leq \rho_{XY} \leq +1$  ist dimensionslos

$\rho_{XY}$  misst die Stärke und Richtung des **monotonen Zusammenhangs** zwischen X und Y.

Der Schätzer des Spearman-Rang-Korrelationskoeffizienten  $\hat{\rho}_{XY}$  entspricht dem des Pearson-Korrelationskoeffizienten  $\hat{r}_{XY}$ , berechnet allerdings für die Ränge der Daten und nicht die Daten selbst. Die besondere Eigenschaft von Rängen erlaubt die Herleitung einer knapperen mathematischen Formel für  $\hat{\rho}_{XY}$  (auf der Folie angegeben), als es für  $\hat{r}_{XY}$  möglich ist.

## Spearman-Rang-Korrelationskoeffizient



Ein exponentieller Zusammenhang ist monoton, aber nicht linear. Daher ist  $r_{XY}$  im vorliegenden Beispiel kleiner als eins, obwohl das exponentielle Regressionsmodell perfekt zu den Daten passt; andererseits gilt aber  $\rho_{XY}=1$ .

Miss America 1984 - 2002

$x_i$	$y_i$	$rg[x_i]$	$rg[y_i]$	$d_i^2$
63"	100 lbs	2	1	1
68"	120 lbs	7	9	4
69"	114 lbs	9.5	5	20.25
68"	116 lbs	7	6	1
70"	131 lbs	11	10	1
63"	108 lbs	2	2	0
68"	118 lbs	7	7.5	0.25
66"	110 lbs	5	3.5	2.25
71"	145 lbs	12	12	0
69"	133 lbs	9.5	11	2.25
64"	118 lbs	4	7.5	12.25
63"	110 lbs	2	3.5	2.25

$$\hat{r}_{XY} = +0.81$$

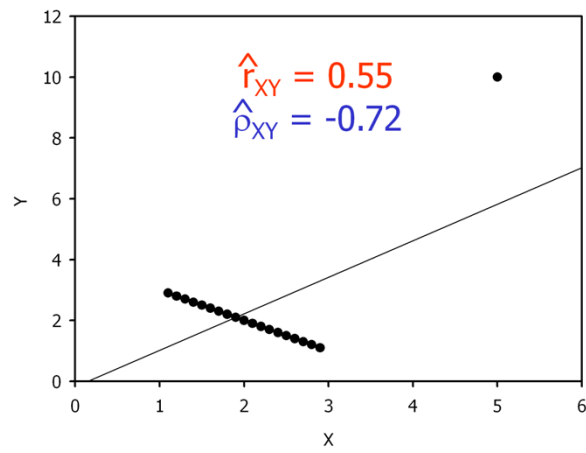
$$\sum_{i=1}^n d_i^2 = 46.5$$

$$\hat{\rho}_{XY} = 0.84$$

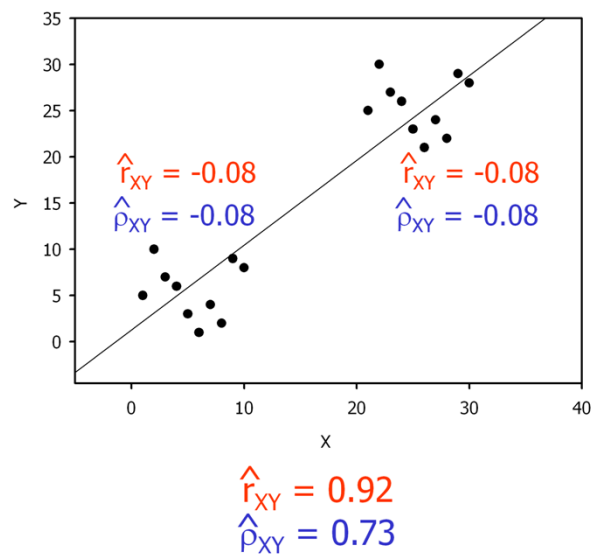
Spearman und Pearson-Korrelationskoeffizient stimmen für die Miss America Daten praktisch überein, was darauf hindeutet, dass das Fehlen eines perfekten linearen Zusammenhangs zwischen Körpergröße und Körpergewicht auf "Rauschen" statt auf eine authentische Nichtlinearität zurückzuführen ist.



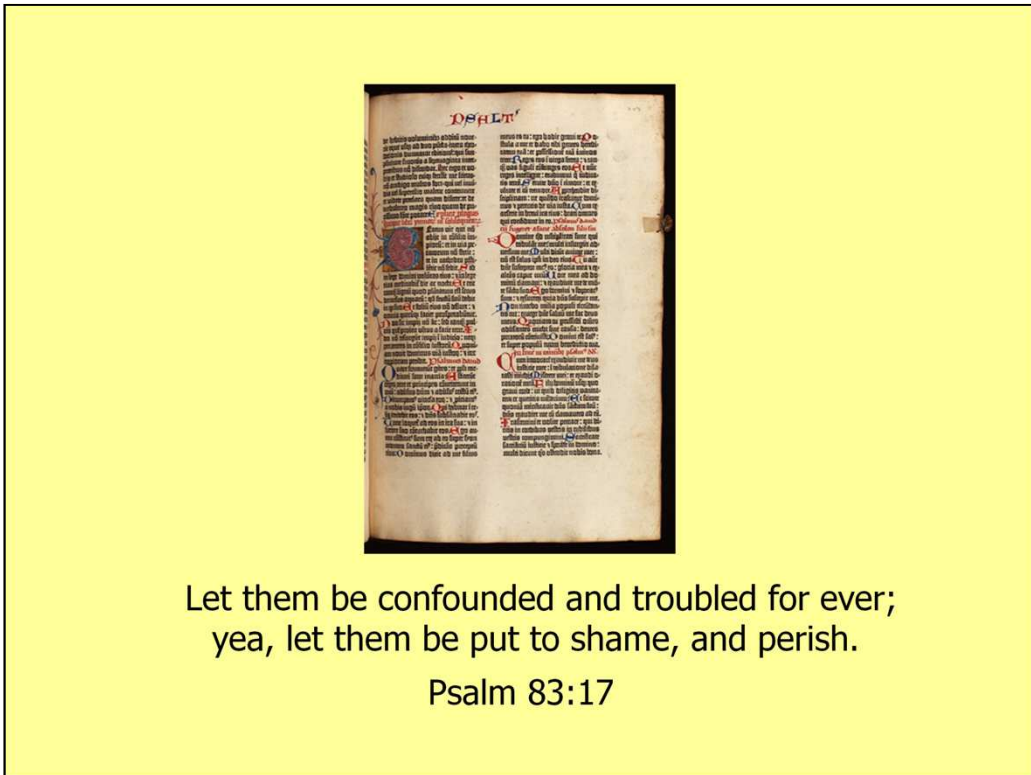
## Spearman-Rang-Korrelationskoeffizient



## Korrelation durch Inhomogenität



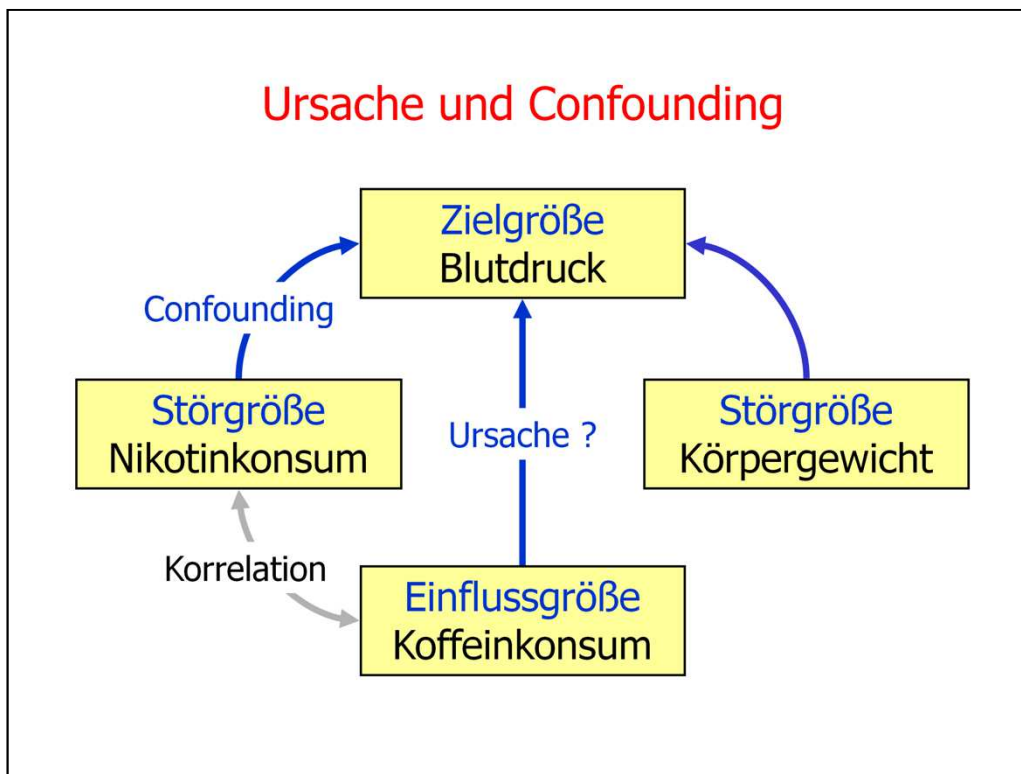
Inhomogenitäten in Daten, aus denen Korrelationskoeffizienten geschätzt werden, stellen eine wesentliche Ursache der Fehleinschätzung von Zusammenhängen dar. Im vorliegenden Beispiel wurden  $\hat{\rho}_{XY}$  und  $\hat{r}_{XY}$  aus Datenpunkten berechnet, die zwei experimentellen Gruppen entstammen. Diese Tatsache wurde bei der Schätzung der Korrelationskoeffizienten jedoch außer Acht gelassen. Die kombinierte Analyse ergab relativ hohe Korrelationskoeffizienten von  $\hat{r}_{XY} = 0.92$  und  $\hat{\rho}_{XY} = 0.73$ , die aber ausschließlich auf das Zusammenfassen beider Gruppen zurückzuführen sind. Sie reflektieren in keiner Weise den "wahren" Zusammenhang zwischen den beiden Variablen, der praktisch gleich null ist. Denken Sie z.B. bei X an die Körpergröße einer Person und bei Y an die tägliche Zeit, die diese Person mit dem Anschauen von Fußballspielen im Fernsehen verbringt. In diesem Fall bestünde die Inhomogenität der Daten darin, dass Frauen und Männer in der Analyse zusammengefasst wurden.



Let them be confounded and troubled for ever;  
yea, let them be put to shame, and perish.

Psalm 83:17

Laut Merriam-Webster Online-Lexikon hat das englische Wort "to confound" mindestens sechs verschiedene Bedeutungen. Eine von ihnen ist "zerstören" oder "ruinieren", und so ist der Gebrauch des Wortes in dem hier gezeigten englischsprachigen Zitat aus Psalm 83 zu interpretieren. Der 83. Psalm ist in der Tat eine ziemlich eindringliche, an Gott gerichtete Bitte, sein Volk beim Kampf gegen dessen Feinde zu unterstützen. Die etymologische Wurzel von "to confound" ist jedoch das lateinische Wort *confundere*, was "vermischen" im Sinne von "verwirren" bedeutet, und in diesem Sinn wird "to confound" im statistischen Kontext verwendet.



Selbst wenn zwei Zufallsvariable zweifelsfrei korreliert sind, bedeutet dies noch keinen notwendigerweise kausalen Zusammenhang zwischen ihnen. Stattdessen kann die Zielgröße stark von einer Störgröße beeinflusst werden, die der Wissenschaftler bislang außer Acht gelassen hat. Wenn diese Variable selbst mit der vermeintlichen Einflussgröße korreliert ist, wirkt sie als "Confounder" der Beziehung zwischen Einfluss- und Zielgröße. Beachten Sie, dass die Bezeichnung "Confounder" kontextabhängig ist. Wenn sich in Abwandlung des vorliegenden Beispiels die wissenschaftliche Fragestellung auf die mögliche Ursächlichkeit des Nikotins selbst bezieht, so würden beide Substanzen die Rollen tauschen, und der Kaffeekonsum wäre als Confounder des Zusammenhangs zwischen Nikotinkonsums und Blutdruck zu betrachten.

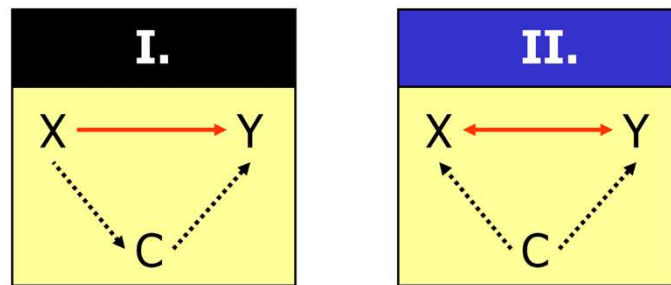
Das Wirken eines Confounders kann in bestimmten Fällen zur Über- oder Unterschätzung des wahren kausalen Zusammenhangs führen. In anderen Situationen kann es sich bei der beobachteten Korrelation zwischen Ziel- und Einflussgröße um ein reines Nebenprodukt der gleichzeitigen Assoziation mit dem Confounder handeln. Das bedeutet insgesamt, dass observationale Daten, die in der Regel nur Korrelationen aufzeigen können, keinen Beweis kausaler Zusammenhänge liefern!

## Confounding

Eine Störgröße wird als **Confounder** des Zusammenhangs zwischen Einfluss- und Zielgröße bezeichnet, wenn sie

- C1: entweder die Zielgröße **beeinflusst** oder das Surrogat einer Einflussgröße ist,
- C2: mit der Einflussgröße **korreliert ist**, und
- C3: keinen **Zwischenschritt** in der Kausalkette zwischen Einfluss- und Zielgröße darstellt.

## Confounding



X: Geschlecht, Y: Verhalten, C: Bildung

X: Therapie, Y: Morbidität, C: Mobilität

X: Ernährung, Y: Lebenserwartung, C: Sozialisation

X: Mobiltelefon, Y: Schlafstörungen, C: Lebensstil

Hinsichtlich seines Einflusses auf die Korrelation zwischen einer Einflussgröße X und einer Zielgröße Y lässt sich der Zusammenhang zwischen X und einem Confounder C irgendwo zwischen zwei Extremen ansiedeln.

Im ersten Szenario (I) übt X sowohl auf C als auch auf Y einen Einfluss aus, aber C beeinflusst Y noch zusätzlich, unabhängig von X. Verhalten (Y) und Bildung (C) hängen z.B. beide sehr stark vom Geschlecht (X) ab (und nicht umgekehrt). Die Bildung trägt aber auch in geschlechtsunabhängiger Weise zum Verhalten eines Kindes bei, was nach Ansicht vieler Psychologen und Soziologen zu einer Überbewertung des Zusammenhangs zwischen Geschlecht und Verhalten führt.

Am anderen Ende der Skala liegt die Situation (II), in der C sowohl X als auch Y beeinflusst (und nicht umgekehrt), so dass die Korrelation zwischen X und Y in keinsten Weise kausal ist. Beim Aufkommen der ersten Mobiltelefone war deren Benutzung (X) zunächst auf Menschen beschränkt, die auf eine steile berufliche Karriere (C) zurückschauen konnten. Ihre Schlafprobleme (Y) waren daher mit großer Sicherheit mehr auf ihren Lebensstil zurückzuführen, als auf die gleichzeitige Benutzung eines Strahlen erzeugenden Kommunikationsmittels.

## Koffein-Abusus (KA) und koronare Herzerkrankung (CHD)

Fall-Kontroll-Studie (144 Patienten, 456 Kontrollen)

KA	CHD	∅CHD
ja	85	215
nein	59	241

$\widehat{OR}=1.6$

*Raucher* →

KA	CHD	∅CHD
ja	80	120
nein	20	20

$\widehat{OR}=0.7$

*Nichtraucher* →

KA	CHD	∅CHD
ja	5	95
nein	39	221

$\widehat{OR}=0.3$

In einer repräsentativen Stichprobe von 144 Patienten und 456 Kontrollen wurde der Zusammenhang zwischen übermäßigem Kaffeekonsum und dem Vorliegen einer koronaren Herzerkrankung (CHD) untersucht. Da die Häufigkeit eines Koffein-Abusus (KA) unter den Patienten (59.0%) deutlich höher war als unter den Kontrollen (47.1%), ergab sich eine geschätzte Odds Ratio von  $(85/59)/(215/241)=1.6$ . Koffeinmissbrauch wäre demnach ein nicht unwesentlicher Risikofaktor für CHD.

Allerdings war KA in der Stichprobe - bei einer geschätzten Odds Ratio von  $(200/40)/(100/260)=13.0$  - extrem stark mit dem Rauchen assoziiert. Da Rauchen selbst ein klassischer Risikofaktor für CHD ist, stellt es in der Studie einen möglichen Confounder des Zusammenhangs zwischen KA und CHD dar.

In der Tat ergibt eine Stratifizierung der Stichprobe nach Rauchverhalten, dass KA in jeder der beiden Subgruppen ein protektiver Faktor für CHD zu sein scheint.

## Kausalitätskriterien



Sir Austin Bradford Hill  
(1897-1991)

"Wissenschaftliches Arbeiten bleibt immer unvollständig - egal ob observational oder experimentell. Jede wissenschaftliche Arbeit läuft Gefahr, durch fortschreitendes Wissen überholt oder überflüssig zu werden. Dies erlaubt es uns aber nicht, derzeitiges Wissen zu ignorieren oder Maßnahmen, die heute ergriffen werden müssen, in die Zukunft zu verschieben."

Austin Bradford Hill, "The Environment and Disease: Association or Causation?,"  
Proceedings of the Royal Society of Medicine, 58  
(1965), 295-300.

Im Jahr 1965 verfasste der britische Medizinstatistiker Sir Austin Bradford Hill einen Artikel, der viele wichtige Anregungen für die Interpretation empirischer Daten enthielt und zu einem Meilenstein der Epidemiologie wurde. Ein wesentliches Augenmerk der Arbeit von Hill galt dem Problem, wie sich auf der Grundlage observationaler Daten ein kausaler Zusammenhang zwischen einer bestimmten Einflussgröße und einer Krankheit herstellen ließe. Man kann zwar einfach behaupten, Einfluss A (z.B. Rauchen) verursache Krankheit B (Lungenkrebs), der wissenschaftliche Nachweis einer solchen Kausalität ist jedoch ungleich schwieriger. Hill schlug daher eine Liste von Kriterien vor, anhand derer sich beurteilen lässt, inwieweit eine beobachtete Korrelation auf einem kausalen Zusammenhang beruht. Ursprünglich für den Einsatz in der Epidemiologie gedacht, haben sich die Bradford-Hill-Kriterien in der Zwischenzeit zu einem grundsätzlichen Instrument der Kausalitätsbewertung in der wissenschaftlichen Forschung entwickelt.

Die Bradford-Hill-Kriterien sind oft dafür kritisiert worden, dass sie für den Nachweis einer Kausalität weder hinreichend noch notwendig sind. Obwohl diese Kritik zweifellos berechtigt ist, so geht sie doch am Kern der Originalarbeit vorbei. Hill war durchaus bewusst, dass ein einzelnes Kriterium für sich genommen nur wenig über die Vertrauenswürdigkeit einer wissenschaftlichen Studie aussagen kann. Dies kann nur das Zusammentreffen vieler verschiedener Faktoren leisten, die im konkreten Fall wissenschaftliche Schlussfolgerungen entweder stützen oder bei Abwesenheit zu Zweifeln an den Aussagen einer Studie berechtigen.



## Kausalitätskriterien

### Stärke (1)

stärkerer Zusammenhang = wahrscheinlichere Kausalität

### Konsistenz (2)

wiederholbar in unterschiedlichen Kontexten

### Spezifität (3)

eine einzelne Ursache für eine gegebene Wirkung

### Temporalität (4)

Ursache geht Wirkung voran

### Dosis-Wirkungs-Beziehung (5)

stärkere Exposition erhöht Wirkung oder Risiko

## Kausalitätskriterien

Plausibilität (6)

mit derzeitigem Wissen erklärbar

Kohärenz (7)

vereinbar mit aktuellen Fakten und Theorien

Experiment (8)

Wirkung ist experimentell steuerbar

Analogie (9)

ähnliche Ursachen für ähnliche Wirkungen

## Zusammenfassung

- Die Stärke und Richtung der Assoziation zwischen zwei quantitativen Zufallsvariablen wird durch ihre **Kovarianz** gemessen.
- Der **Pearson-Korrelationskoeffizient** misst die Stärke und Richtung des linearen Zusammenhangs zwischen zwei quantitativen Zufallsvariablen.
- **Regressionsanalysen** erzeugen mathematische Modelle für den quantitativen Zusammenhang zwischen Zufallsvariablen.
- **Kausalität** folgt nicht notwendigerweise aus dem Vorliegen einer Korrelation, sondern bedarf zusätzlicher Evidenz (z.B. gemäß Bradford-Hill-Kriterien).